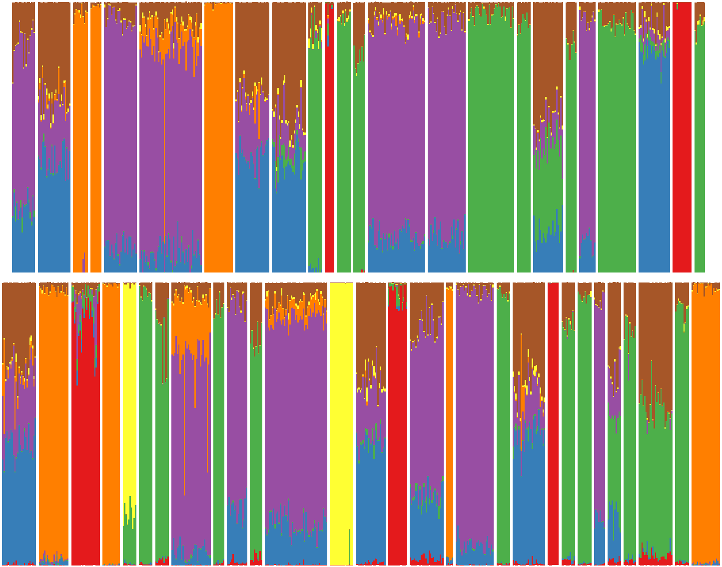
MODERN PROBABILISTIC MODELING FOR MASSIVE DATA

David M. Blei
Columbia University

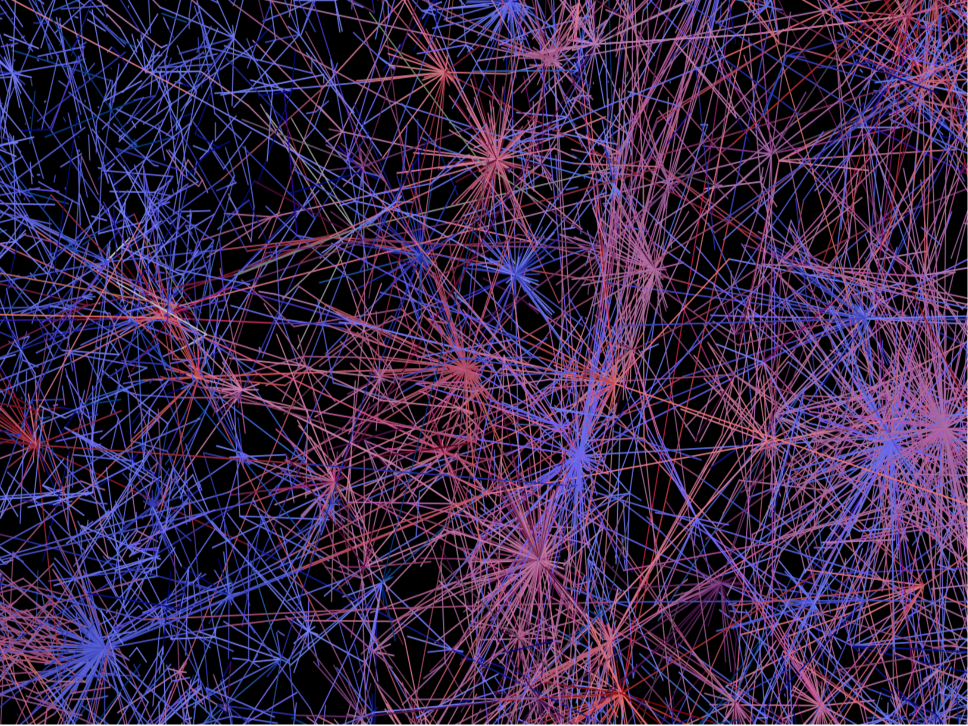
Modern probabilistic modeling:

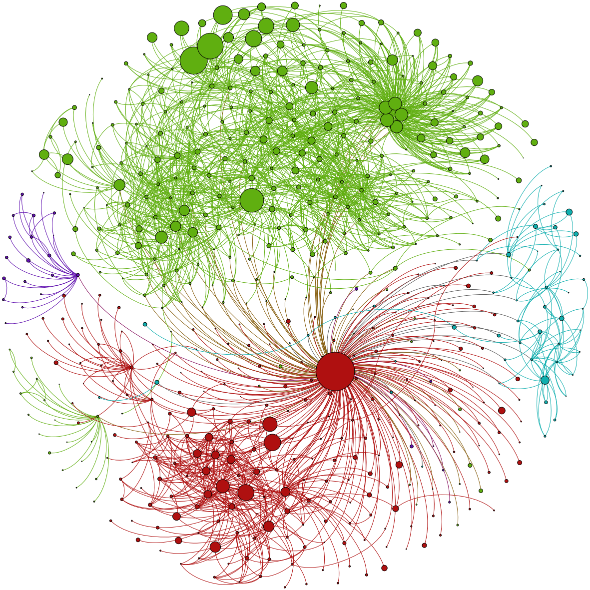
An efficient framework for discovering useful patterns in massive data.

[Amy Pettingill]



Population analysis of 2 billion genetic measurements





Communities discovered in a 3.7M node network of U.S. Patents

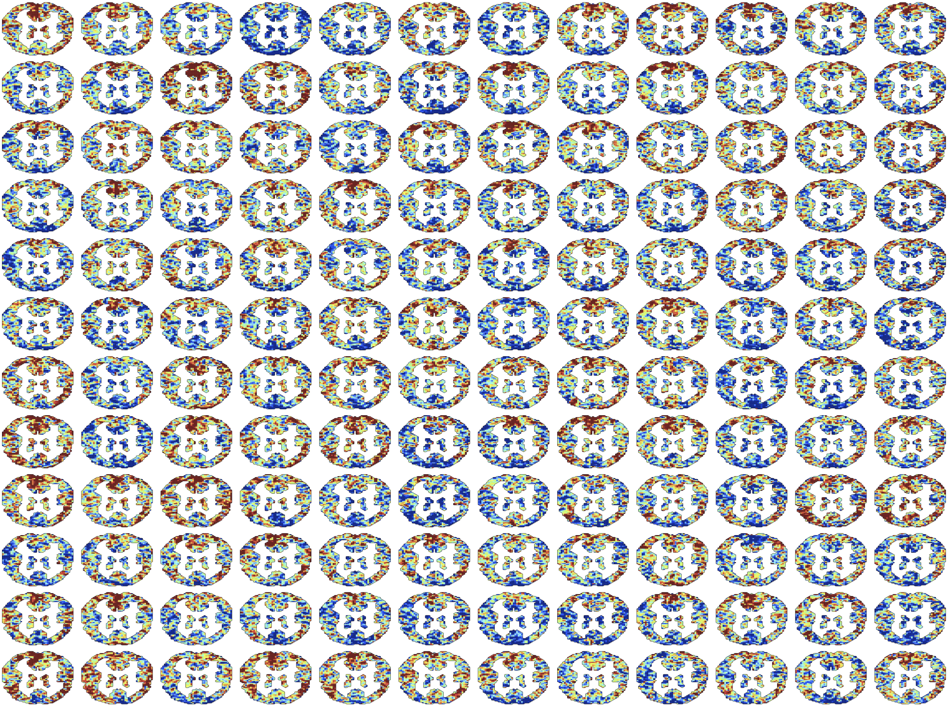

The New York Times
NEW YORK ALLOWS SAME-SEX MARRIAGE, BECOMING LARGEST STATE TO PASS LAW
International New York Times
May 10, 1943
WARSAW GHETTO UPRISING AN OVER-REACTION
EUROPEAN LEADERS BLAME JEWS FOR DISPROPORTIONATE RESPONSE
Front Page News
LINDBERGH DOES IT! TO PARIS IN 33½ HOURS; FLIES 1,000 MILES THROUGH SNOW AND SLEET; CHEERING FRENCH CARRY HIM OFF FIELD
127 DIE AS 2 AIRLINERS COLLIDE OVER CITY; JET SETS BROOKLYN FIRE, KILLING 5 OTHERS; SECOND PLANE CRASHES ON STATEN ISLAND
THE SHUTTLE EXPLODES
6 IN CREW AND HIGH-SCHOOL TEACHER ARE KILLED 74 SECONDS AFTER LIFTOFF
As OSHA Emphasizes Safety, Long-Term Health Risks Fester
Mission Accomplished!
OBAMA WINS!!!
Sulzberger says, Pinch me, I must be dreaming!
Obama thanks New York Times for media support, says, Yes we did!
Behind TV Analysts, Pentagon's Hidden Hand
NIXON RESIGNS
The Washington Post
Cardinal Wojtyla of Poland Is Elected Pope

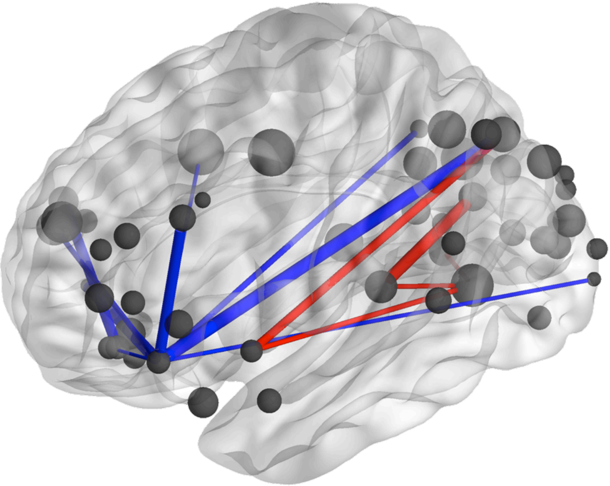
1	2	3	4	5
Game	Life	Film	Book	Wine
Season	Know	Movie	Life	Street
Team	School	Show	Books	Hotel
Coach	Street	Life	Novel	House
Play	Man	Television	Story	Room
Points	Family	Films	Man	Night
Games	Says	Director	Author	Place
Giants	House	Man	House	Restaurant
Second	Children	Story	War	Park
Players	Night	Says	Children	Garden

6	7	8	9	10
Bush	Building	Won	Yankees	Government
Campaign	Street	Team	Game	War
Clinton	Square	Second	Mets	Military
Republican	Housing	Race	Season	Officials
House	House	Round	Run	Iraq
Party	Buildings	Cup	League	Forces
Democratic	Development	Open	Baseball	Iraqi
Political	Space	Game	Team	Army
Democrats	Percent	Play	Games	Troops
Senator	Real	Win	Hit	Soldiers

11	12	13	14	15
Children	Stock	Church	Art	Police
School	Percent	War	Museum	Yesterday
Women	Companies	Women	Show	Man
Family	Fund	Life	Gallery	Officer
Parents	Market	Black	Works	Officers
Child	Bank	Political	Artists	Case
Life	Investors	Catholic	Street	Found
Says	Funds	Government	Artist	Charged
Help	Financial	Jewish	Paintings	Street
Mother	Business	Pope	Exhibition	Shot

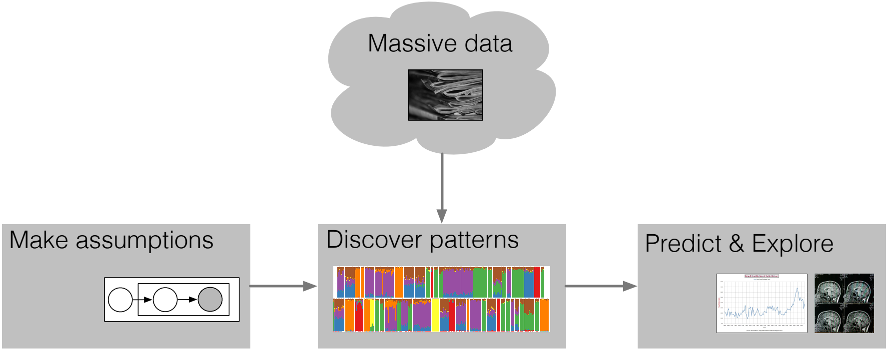
Topics found in 1.8M articles from the New York Times





Neuroscience analysis of 220 million fMRI measurements

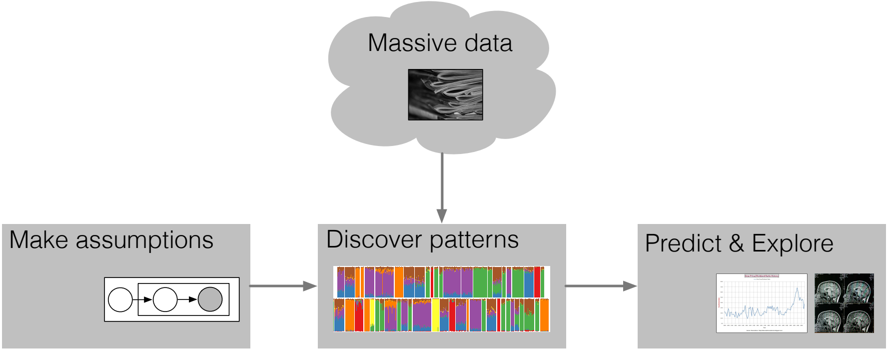
The probabilistic modeling pipeline



Our perspective:

- This is a framework for **customized data analysis**, crucial to many fields.
- The pipeline separates assumptions, computation, application
- It facilitates solving modern data science problems.

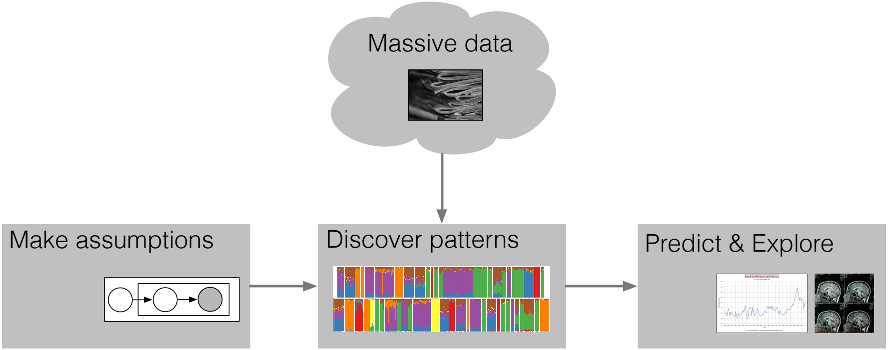
The probabilistic modeling pipeline



Our goal:

- Develop modeling into a **flexible**, **powerful** and **easy-to-use** way to solve real-world problems.

The probabilistic modeling pipeline



Our challenges:

- Develop new ways to build **flexible models**
- Develop algorithms that work on many problems and with **massive data**.
- Solve **new problems** in science, industry, and government

Jaan Aaltosar

Allison Chaney

Rajesh Ranganath

Maja Rudolph

Laurent Charlin

Alp Kucukelbir

Stephan Mandt

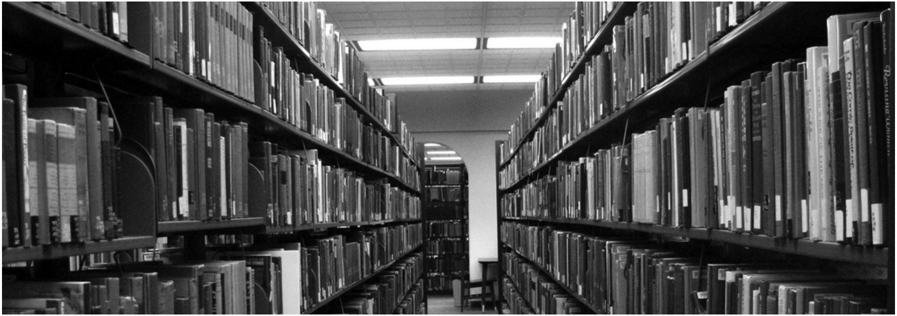
Jeremy Manning

James McInerney

Probabilistic topic models

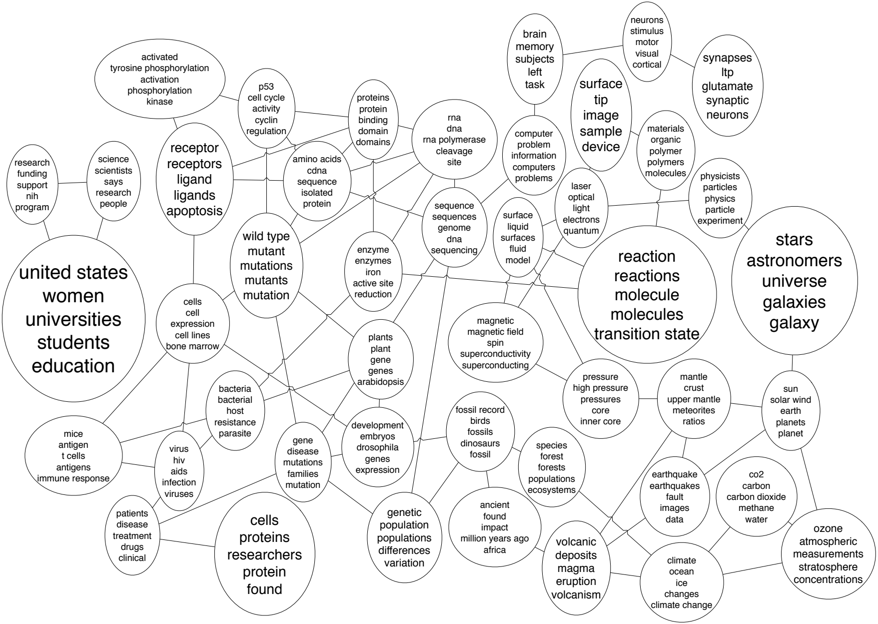
Powerful and flexible algorithms for analyzing massive collections of text

- ORGANIZE
- VISUALIZE
- SUMMARIZE
- SEARCH
- PREDICT
- UNDERSTAND



TOPIC MODELING

1. **Discover** the thematic structure
2. **Annotate** the documents
3. **Use** the annotations to visualize, organize, summarize, ...



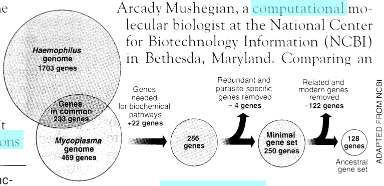
activated tyrosine phosphorylation activation phosphorylation kinase
p53 cell cycle activity cyclin regulation
proteins protein binding domain domains
rna dna rna polymerase cleavage site
brain memory subjects left task
neurons stimulus motor visual cortical
synapses ltp glutamate synaptic neurons
surface tip image sample device
materials organic polymer polymers molecules
computer problem information computers problems
research funding support nih program
science scientists says research people
receptor receptors ligand ligands apoptosis
amino acids cdna sequence isolated protein
laser optical light electrons quantum
physicists particles physics particle experiment
sequence sequences genome dna sequencing
surface liquid surfaces fluid model
wild type mutant mutations mutants mutation
enzyme enzymes iron active site reduction
reaction reactions molecule molecules transition state
stars astronomers universe galaxies galaxy
united states women universities students education
cells cell expression cell lines bone marrow
magnetic magnetic field spin superconductivity superconducting
plants plant gene genes arabidopsis
pressure high pressure pressures core inner core
mantle crust upper mantle meteorites ratios
sun solar wind earth planets planet
bacteria bacterial host resistance parasite
fossil record birds fossils dinosaurs fossil
development embryos drosophila genes expression
species forest forests populations ecosystems
mice antigen t cells antigens immune response
virus hiv aids infection viruses
gene disease mutations families mutation
earthquake earthquakes fault images data
co2 carbon carbon dioxide methane water
ancient found impact million years ago africa
ozone atmospheric measurements stratosphere concentrations
patients disease treatment drugs clinical
cells proteins researchers protein found
genetic population populations differences variation
volcanic deposits magma eruption volcanism
climate ocean ice changes climate change

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

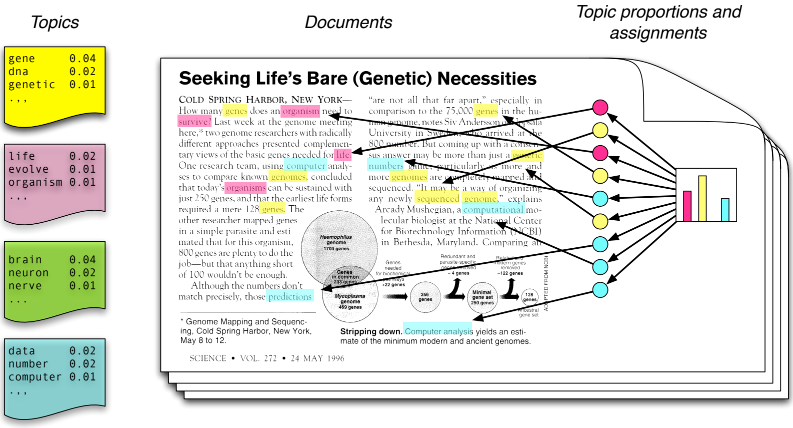
Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

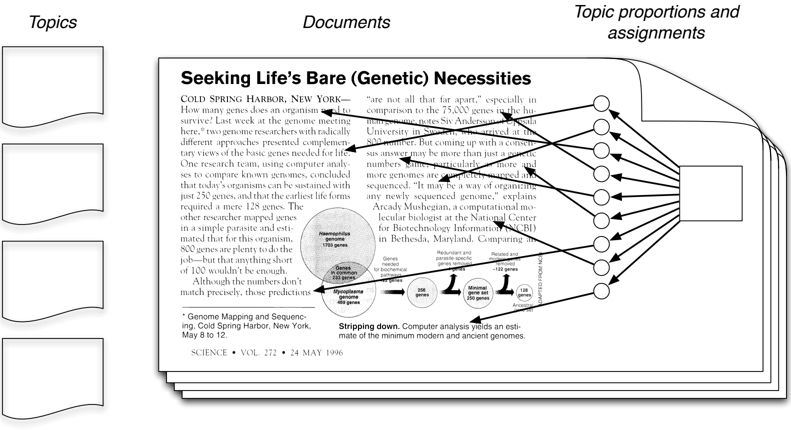


Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

Documents exhibit multiple topics.



Latent Dirichlet Allocation



Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

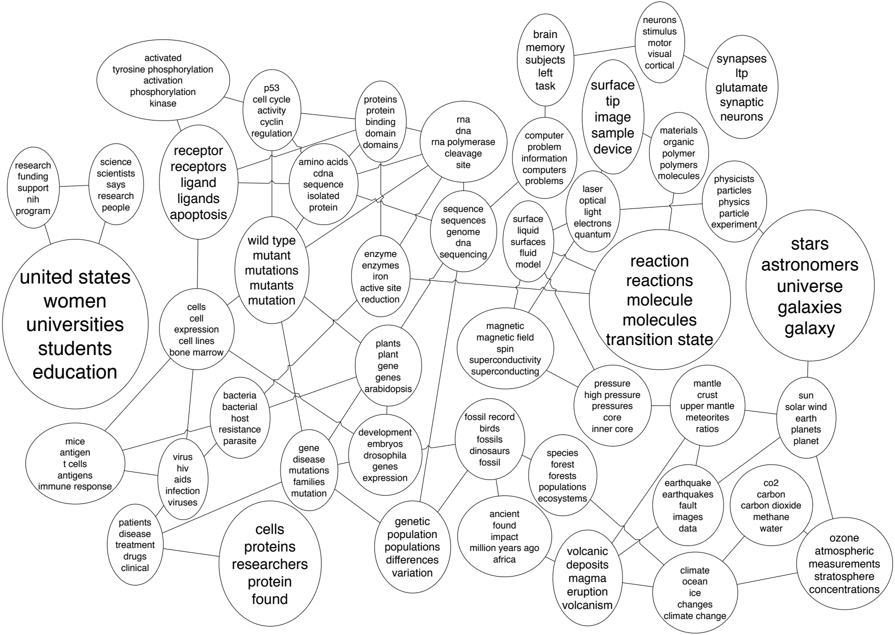
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

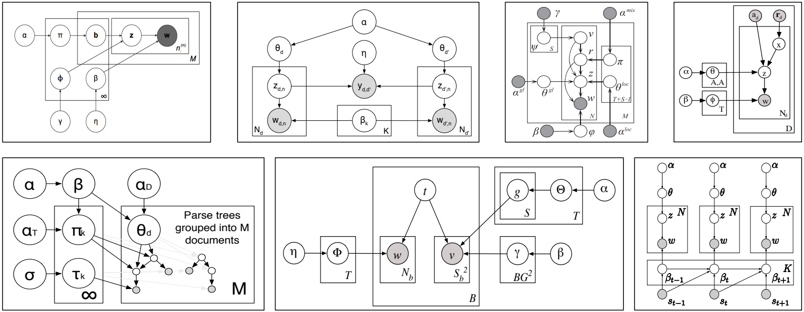
SCIENCE • VOL. 272 • 24 MAY 1996

Latent Dirichlet Allocation

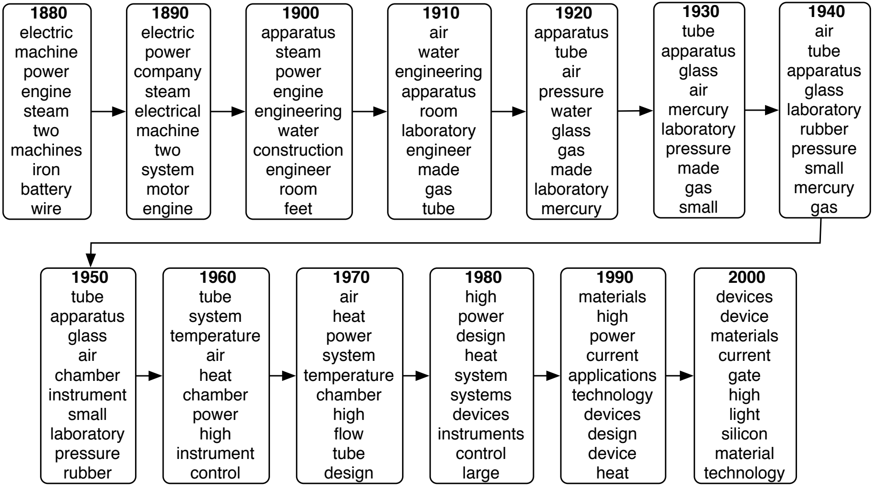
- **Data**: The OCR'ed collection of *Science* from 1990–2000
  - 17K documents
  - 11M words
  - 20K unique terms (stop words and rare words removed)
- **Model**: 100-topic LDA model using variational inference.


activated tyrosine phosphorylation activation phosphorylation kinase
p53 cell cycle activity cyclin regulation
proteins protein binding domain domains
rna dna rna polymerase cleavage site
brain memory subjects left task
neurons stimulus motor visual cortical
synapses ltp glutamate synaptic neurons
surface tip image sample device
materials organic polymer polymers molecules
computer problem information computers problems
research funding support nih program
science scientists says research people
receptor receptors ligand ligands apoptosis
amino acids cdna sequence isolated protein
physicists particles physics particle experiment
laser optical light electrons quantum
surface liquid surfaces fluid model
sequence sequences genome dna sequencing
wild type mutant mutations mutants mutation
enzyme enzymes iron active site reduction
reaction reactions molecule molecules transition state
stars astronomers universe galaxies galaxy
united states women universities students education
cells cell expression cell lines bone marrow
magnetic magnetic field spin superconductivity superconducting
plants plant gene genes arabidopsis
pressure high pressure pressures core inner core
mantle crust upper mantle meteorites ratios
sun solar wind earth planets planet
bacteria bacterial host resistance parasite
development embryos drosophila genes expression
fossil record birds fossils dinosaurs fossil
species forest forests populations ecosystems
mice antigen t cells antigens immune response
virus hiv aids infection viruses
gene disease mutations families mutation
earthquake earthquakes fault images data
co2 carbon carbon dioxide methane water
ozone atmospheric measurements stratosphere concentrations
patients disease treatment drugs clinical
cells proteins researchers protein found
genetic population populations differences variation
ancient found impact million years ago africa
volcanic deposits magma eruption volcanism
climate ocean ice changes climate change

- LDA builds on decades of research about how to **derive meaning from text**.
- LDA more easily **scales to massive data** and **generalizes to new data**.
- LDA has had a big impact on many fields
  - Natural language processing
  - Computer vision
  - Recommendation systems
  - Web search
  - Computational biology and genetics

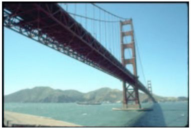


- LDA is a simple **building block** that enables many applications.
- Each model solves a different problem, fuses different kinds of data.
- Models and their algorithms easily compose.


1880
electric
machine
power
engine
steam
two
machines
iron
battery
wire
1890
electric
power
company
steam
electrical
machine
two
system
motor
engine
1900
apparatus
steam
power
engine
engineering
water
construction
engineer
room
feet
1910
air
water
engineering
apparatus
room
laboratory
engineer
made
gas
tube
1920
apparatus
tube
air
pressure
water
glass
gas
made
laboratory
mercury
1930
tube
apparatus
glass
air
mercury
laboratory
pressure
made
gas
small
1940
air
tube
apparatus
glass
laboratory
rubber
pressure
small
mercury
gas
1950
tube
apparatus
glass
air
chamber
instrument
small
laboratory
pressure
rubber
1960
tube
system
temperature
air
heat
chamber
power
high
instrument
control
1970
air
heat
power
system
temperature
chamber
high
flow
tube
design
1980
high
power
design
heat
system
systems
devices
instruments
control
large
1990
materials
high
power
current
applications
technology
devices
design
device
heat
2000
devices
device
materials
current
gate
high
light
silicon
material
technology

SKY WATER TREE MOUNTAIN PEOPLE

SCOTLAND WATER FLOWER HILLS TREE



SKY WATER BUILDING PEOPLE WATER

FISH WATER OCEAN TREE CORAL

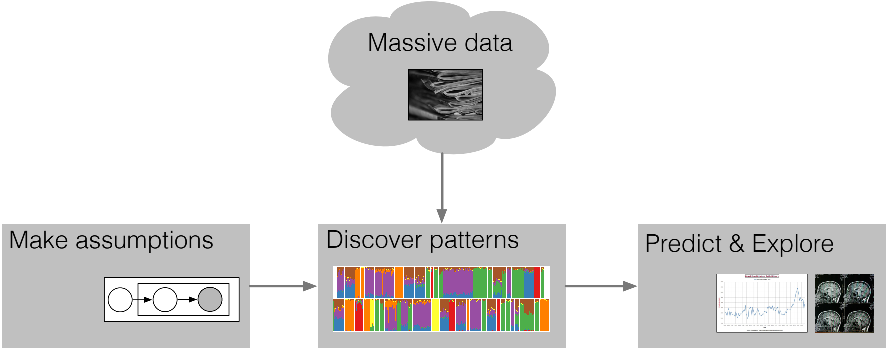
PEOPLE MARKET PATTERN TEXTILE DISPLAY

BIRDS NEST TREE BRANCH LEAVES

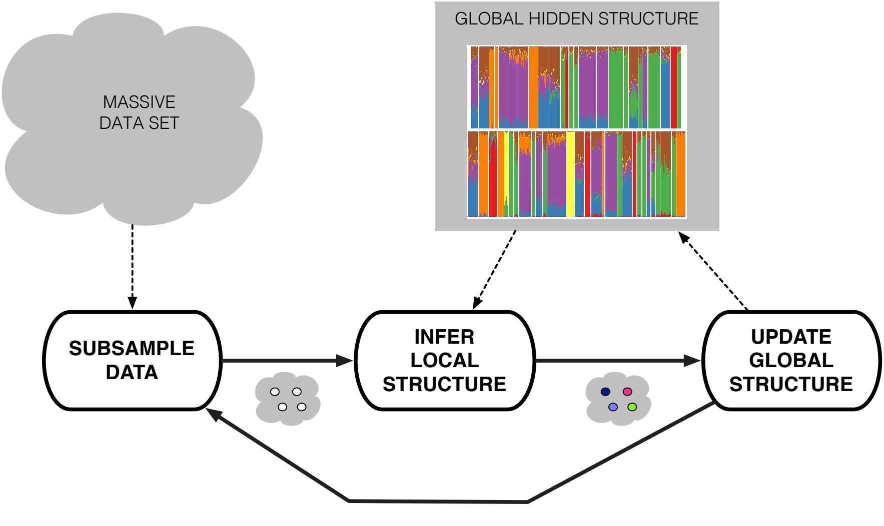

Heard a loud bang nearby, what was that?
12:11pm
Something flew into the ground and exploded
12:14pm
Oh my god, just saw a helicopter crash
12:16pm
Barneys New York
Lexington Av/63 St [F]
Apple Store
FAO Schwarz
Lexington Av/ 59 St [N,Q,R]
59 St [4,5,6]
Bloomingdale's
Dylan's Candy Bar
Bloomberg LP
Dangerfield
Sofrito
Lexington Av/ 53 St [E,M]
E 63rd St
E 62nd St
E 61st St
E 64th St
E 65th St
66th St
E 58th St
E 56th St
E 55th St
E 54th St
E 53rd St
E 52nd St
E 51st St
E 59th
E 57th St
Madison Ave
Park Ave
3rd Ave
2nd Ave

Probabilistic inference

Given a model, use an algorithm to discover the hidden patterns in the data.



- Probabilistic inference is the main **algorithmic & statistical problem**.
- We square the modeling assumptions with the observed data.
  E.g., which topics likely generated a collection of documents?
- We need **scalable** and **generic** inference.



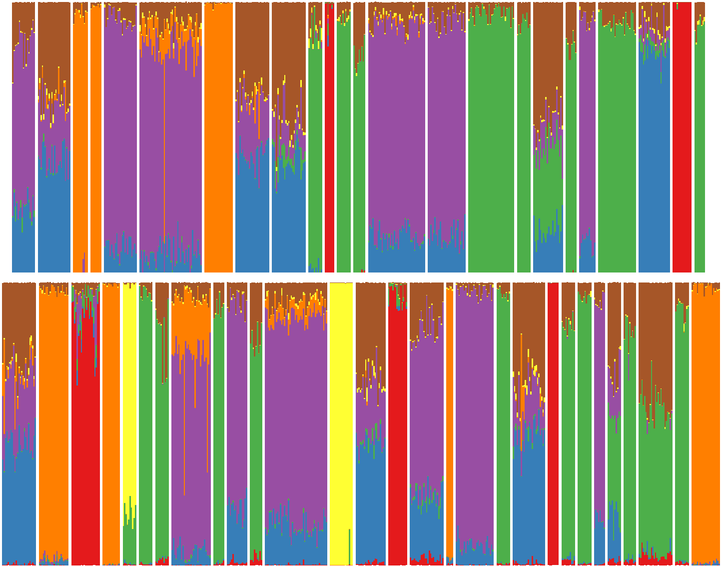
Stochastic variational inference scales to **massive data**.

1	2	3	4	5
Game	Life	Film	Book	Wine
Season	Know	Movie	Life	Street
Team	School	Show	Books	Hotel
Coach	Street	Life	Novel	House
Play	Man	Television	Story	Room
Points	Family	Films	Man	Night
Games	Says	Director	Author	Place
Giants	House	Man	House	Restaurant
Second	Children	Story	War	Park
Players	Night	Says	Children	Garden

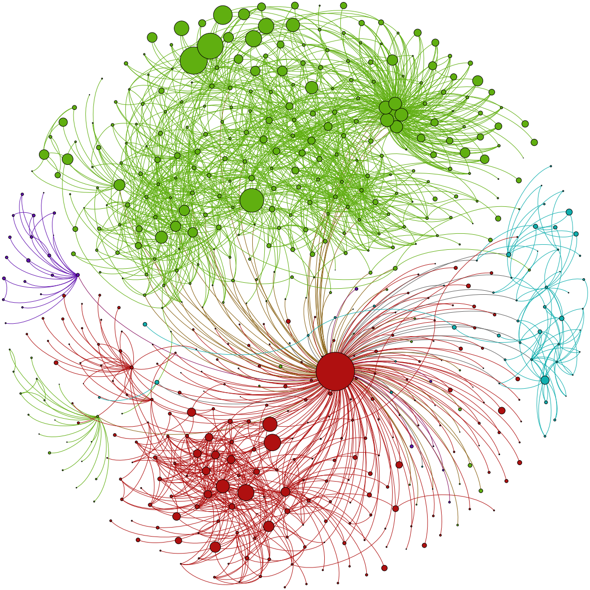
6	7	8	9	10
Bush	Building	Won	Yankees	Government
Campaign	Street	Team	Game	War
Clinton	Square	Second	Mets	Military
Republican	Housing	Race	Season	Officials
House	House	Round	Run	Iraq
Party	Buildings	Cup	League	Forces
Democratic	Development	Open	Baseball	Iraqi
Political	Space	Game	Team	Army
Democrats	Percent	Play	Games	Troops
Senator	Real	Win	Hit	Soldiers

11	12	13	14	15
Children	Stock	Church	Art	Police
School	Percent	War	Museum	Yesterday
Women	Companies	Women	Show	Man
Family	Fund	Life	Gallery	Officer
Parents	Market	Black	Works	Officers
Child	Bank	Political	Artists	Case
Life	Investors	Catholic	Street	Found
Says	Funds	Government	Artist	Charged
Help	Financial	Jewish	Paintings	Street
Mother	Business	Pope	Exhibition	Shot

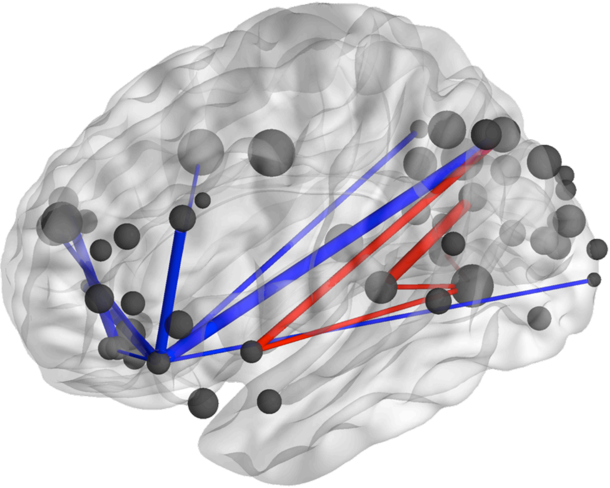
Topics found in 1.8M articles from the New York Times



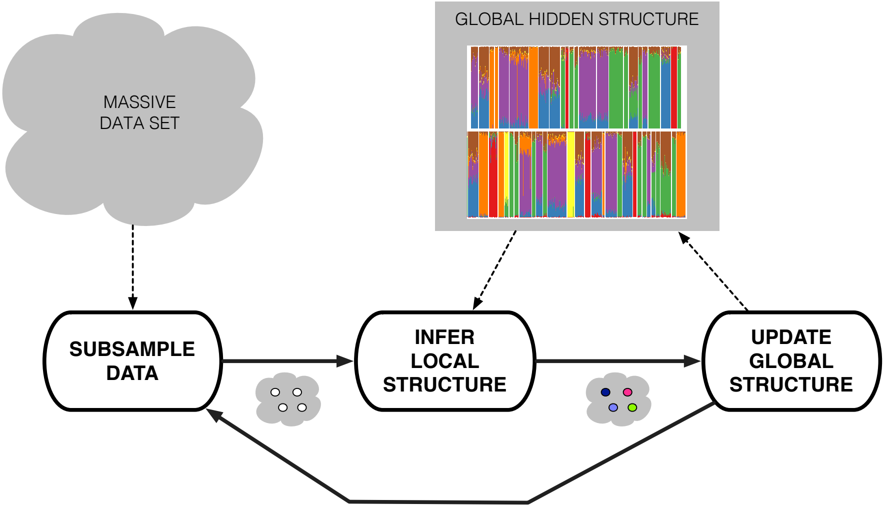
Population analysis of 2 billion genetic measurements



Communities discovered in a 3.7M node network of U.S. Patents



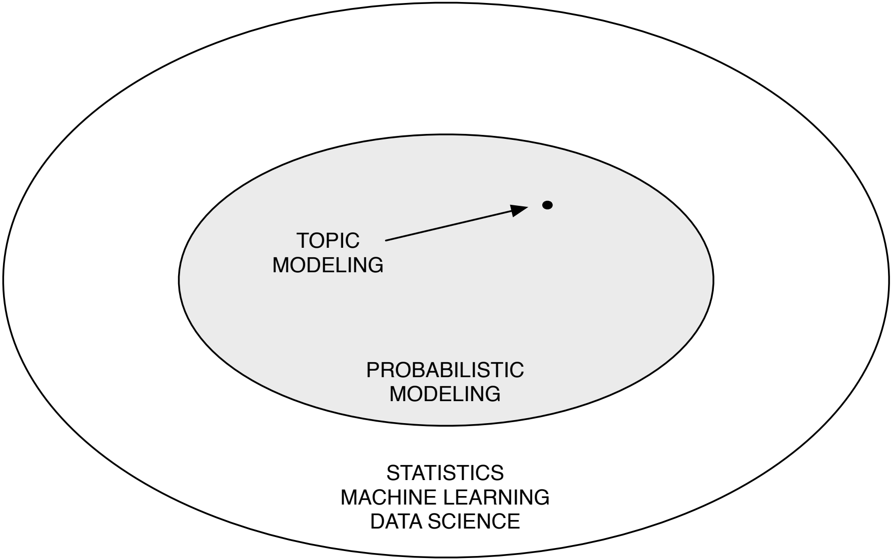
Neuroscience analysis of 220M fMRI measurements



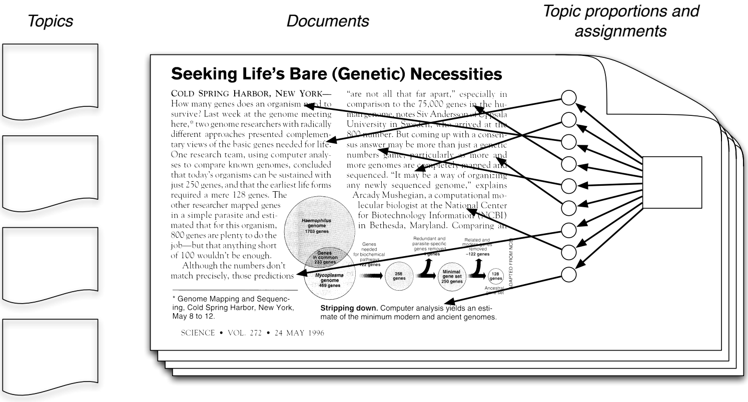
- Uses **stochastic optimization** (Robbins & Monro, 1951)
- Scales up 50 years of research in Bayesian modeling
- Though these are recent results, they have been adapted to many domains

Modern probabilistic modeling:

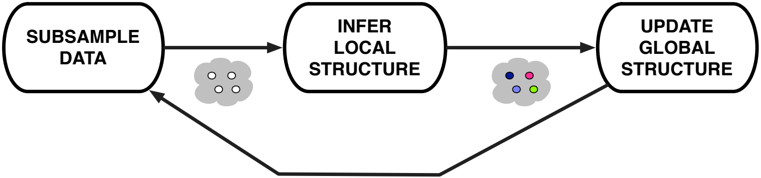
An efficient framework for discovering useful patterns in massive data.


TOPIC
MODELING
PROBABILISTIC
MODELING
STATISTICS
MACHINE LEARNING
DATA SCIENCE

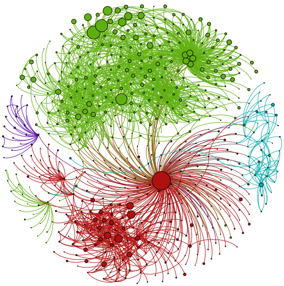
I. Assume our data come from a model with hidden patterns at work

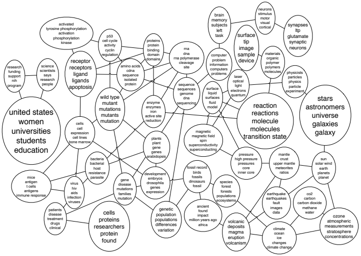


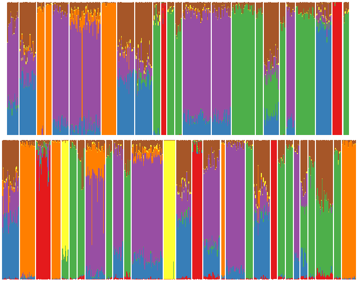
II. Discover those patterns in the data

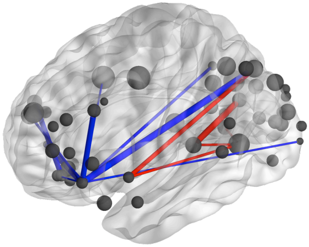


III. Use the discovered patterns to predict about and explore the data

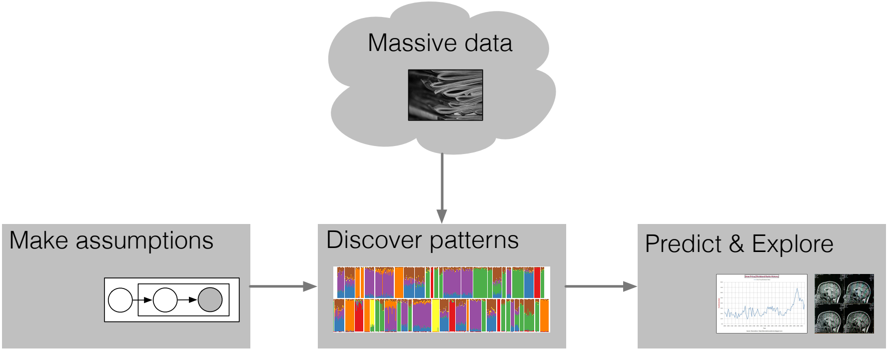






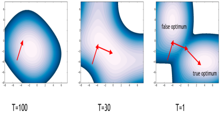


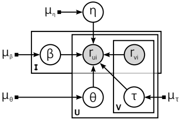
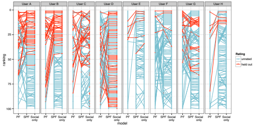
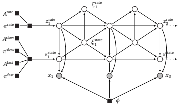
The probabilistic modeling pipeline



Our goal:

- Develop modeling into a **flexible**, **powerful** and **easy-to-use** way to solve real-world problems.



Models and applications

- genetic measurements
- hierarchical topics in a corpus
- changing preferences
- equations and text
- newsworthy events in Twitter
- news consumption in a network
- word meanings
- counselor/patient dialogs
- declassified cables from the 70s
- neural readings in a fish

Inference

- active subsampling
- averaged gradients
- annealing and inference
- stochastic optimization
- structured variational inference
- probabilistic programming

We should seek out unfamiliar summaries of observational material, and establish their useful properties... And still more novelty can come from finding, and evading, still deeper lying constraints.

(John Tukey, *The Future of Data Analysis*, 1962)